Reliability of MRI findings in candidates for lumbar disc prosthesis

Linda Berg · Gesche Neckelmann · Øivind Gjertsen · Christian Hellum · Lars G. Johnsen · Geir E. Eide · Ansgar Espeland

Received: 30 May 2011 / Accepted: 13 September 2011 / Published online: 23 September 2011

Abstract

Introduction Limited reliability data exist for localised magnetic resonance imaging (MRI) findings relevant to planning of treatment with lumbar disc prosthesis and later outcomes. We assessed the reliability of such findings in chronic low back pain patients who were accepted candidates for disc prosthesis.

Methods On pretreatment MRI of 170 patients (mean age 41 years; 88 women), three experienced radiologists independently rated Modic changes, disc findings and facet arthropathy at L3/L4, L4/L5 and L5/S1. Two radiologists rerated 126 examinations. For each MRI finding at each disc level, agreement was analysed using the kappa statistic and differences in prevalence across observers using a fixed effects model.

Results All findings at L3/L4 and facet arthropathy at L5/S1 had a mean prevalence <10% across observers and were not further analysed, ensuring interpretable kappa values. Overall interobserver agreement was generally moderate or good (kappa 0.40–0.77) at L4–S1 for Modic changes, nucleus pulposus signal, disc height (subjective and measured), posterior high-intensity zone (HIZ) and disc contour, and fair (kappa 0.24) at L4/L5 for facet arthropathy. Posterior HIZ at L5/S1 and severely reduced subjective disc height at L4/L5 differed up to threefold in prevalence between observers ($p<$ 0.0001). Intraobserver agreement was mostly good or very good (kappa 0.60–1.00).

Conclusion In candidates for disc prosthesis, mostly moderate interobserver agreement is expected for localised MRI findings.

L. Berg · G. Neckelmann · A. Espeland (✉)
Department of Radiology, Haukeland University Hospital,
Jonas Liesvei 65,
5021 Bergen, Norway
e-mail: ansgar.espeland@helse-bergen.no

L. Berg · A. Espeland
Section for Radiology, Department of Surgical Sciences,
University of Bergen,
Bergen, Norway

Ø. Gjertsen
Department of Neuroradiology, Oslo University Hospital,
Oslo, Norway

C. Hellum
Department of Orthopaedics, Oslo University Hospital,
Oslo, Norway

C. Hellum
Department of Orthopaedics, University of Oslo,
Oslo, Norway

L. G. Johnsen
National Centre for Diseases of the Spine,
University Hospital of Trondheim,
Trondheim, Norway

L. G. Johnsen
Orthopaedic Department, University Hospital of Trondheim,
Trondheim, Norway

G. E. Eide
Centre for Clinical Research, Haukeland University Hospital,
Bergen, Norway

G. E. Eide
Department of Public Health and Primary Health Care,
University of Bergen,
Bergen, Norway

Keywords Degeneration · Disc prosthesis · Lumbar spine · Magnetic resonance imaging · Reliability

Introduction

Lumbar surgery with fusion or disc prosthesis is being evaluated in clinical studies as treatment for patients with chronic low back pain (LBP) [1–3]. Single or two-level disc degeneration on magnetic resonance imaging (MRI) is a proposed part of the indication for such treatment, and adjacent level and facet degeneration are important issues in these patients [3–5]. Reliable assessment of findings from MRI is crucial to decide on and plan the surgery, to assess its effects, and to study the prognostic role of MRI findings. Unreliable findings in clinical practice and research can lead to incorrect treatment, faulty assessment of adjacent level and facet degeneration, and underestimation of the findings' potential relationship to clinical features and prognosis [6, 7].

Adequate agreement on both type and prevalence of MRI findings at individual disc levels is required to study which and how many levels to treat, to assess the prevalence of any later adjacent level degeneration, and to evaluate how the localised findings may affect prognosis. Therefore, we need data not only on observer agreement (kappa values) but also on differences in reported prevalence of relevant MRI findings between observers at separate disc levels.

Previous studies have examined observer agreement for relevant MRI findings, such as Modic changes [8–12], posterior high-intensity zone (HIZ) in the disc [9, 10, 12–15], disc degeneration [9, 10, 12, 15], abnormal disc contour [9, 12, 15, 16] and facet arthropathy [10, 17, 18]. However, differences between observers in the reported prevalence of such findings have received very little attention [10, 16]. Some of the prior studies had only two observers [8, 11, 13–15, 17] and/or a modest sample size [8, 9, 11, 12, 16–18], focused on one or a few findings [8, 13, 17, 18] and/or reported combined results for several disc levels [9, 10]. Only one study concerned disc prosthesis patients, and it was restricted to facet arthropathy [18].

The aim of the present study was to assess the reliability of pretreatment lumbar spine MRI findings in chronic LBP patients who were accepted candidates for lumbar disc prosthesis. At each disc level for each MRI finding, we analysed interobserver and intraobserver agreement as well as differences in reported prevalence among experienced radiologists. Such analyses at individual levels were also done for combined findings used as MRI indication for prosthesis.

Materials and methods

The appropriate regional research ethics committee approved this study. All patients gave their informed consent prior to their inclusion in the study.

Patients

Of 173 LBP patients randomized to disc prosthesis surgery or multidisciplinary rehabilitation in a prospective national trial [3], 170 (98.3%; mean age 41 years; 82 men, 88 women) had pretreatment MRI available for this retrospective reliability study. The results of this study were not used to determine eligibility in the trial and have not been published previously. The criteria for inclusion in the trial were: age 25–55 years, LBP as main symptom for at least 1 year, insufficient effect of physiotherapy or chiropractic treatment, Oswestry Disability Index (ODI) ≥30% and the following MRI findings reported by the enrolling physicians at L4/L5 and/or at L5/S1 (levels suitable for disc prosthesis): (a) ≥40% disc height decrease compared to the nearest normal above disc and/or (b) at least two of these three findings: Modic changes type I (oedema) and/or type II (fat), posterior HIZ in the disc and dark/black nucleus pulposus on T2-weighted images. Patients were excluded if they had any of the four findings in a or b at any higher lumbar level (L1–L4) or had spondylolysis, spondylolisthesis, arthritis, osteoporosis, prior fracture L1–S1, prior spinal fusion, deformity, or symptomatic disc herniation/spinal stenosis. Facet joint degeneration was not an exclusion criterion.

Images

MRI was performed as part of clinical practice, using different protocols and magnets (1.5 T in 150 of 170 cases). All examinations included sagittal T2-weighted fast spin echo images: repetition time (TR)/echo time (TE), 2,511–4,760 ms/91–140 ms. All but two (168/170) included sagittal T1-weighted images: 159 spin echo images (TR/TE, 350–91 ms/7–22 ms) and 9 T1 fast fluid-attenuated inversion-recovery images (TR/TE, 1,984–2,130 ms/20–22 ms). Most (168/170) included axial images of the L4/L5 and L5/S1 levels: 135 T2-, 33 T1- and 21 proton density-weighted images. Few (5/170) included sagittal fat-suppression images. Typically, slice thickness was 3–5 mm, interslice gap 0.3–2.2 mm, field of view 19–38 cm for sagittal and 15–32 cm for axial images, and matrix 512×512 in the sagittal (115/170) and in the axial plane (89/170). Matrix varied from 160×256 to 640×640. The images were obtained directly in DICOM format or, in seven cases, as digitized printed film hard copies stored in DICOM format and were de-identified before being evaluated.

Ratings

One radiologist experienced in musculoskeletal MRI (A) and two neuroradiologists (B and C) from three different institutions rated findings on the images. Each observer had more than 10 years experience in reporting lumbar spine MRI findings. Observers A and C viewed the images on a clinical PACS unit and observer B on a personal computer. Observers A and B used the eFilm Lite software version 2.1.2 (Merge Healthcare, Hartland, Wisconsin), while observer C used the Agfa Impax 4.5 (Agfa HealthCare, Mortsel, Belgia).

We used existing MRI rating criteria for Modic changes [11, 19–21], posterior HIZ in the disc [10, 14], nucleus pulposus signal [22], disc height (subjective and measured) [15, 23–25], disc contour [19] and facet arthropathy [10, 26] (Table 1). Facet arthropathy was rated using Fujiwara and colleagues' simple system [26] combined with illustrations from the Spine Pain Outcomes Research Trial, which had yielded better agreement than Weishaupt and colleagues' system [10]. The observers also received published illustrations of Modic changes and HIZ [10]. They selected ratings from multiple choice lists for each variable at each of the disc levels L3/L4, L4/L5 and L5/S1. The types (none, I, II, III; primary and secondary), anteroposterior (AP) extent, and craniocaudal (CC) extent of Modic changes were rated both inferiorly and superiorly to the disc. Ratings were dichotomized as shown in the "Results" section prior to the statistical analysis.

Blinded to clinical data and each others' ratings, all three observers evaluated the 170 MRI examinations in random order over 3–4 months. They were asked to also rate the variables on images of suboptimal quality, since these images had been accepted on enrolment and reflected practice. Blinded to and >3 months after their first rating, two observers (A and B) rerated 126 examinations in a new random order. These examinations were selected because the reratings were needed for comparison purposes in a follow-up study of these patients, who were also imaged at the end of 2 years of follow-up. These 126 patients were similar to the rest (n=44) of the 170 patients in gender (p=0.938; chi-squared test) and ODI (p=0.278; t test, normal distribution) and were only slightly older (mean age 41.6 vs. 38.9 years in the n=44 group; p=0.027; t test, normal distribution).

Pilot study

To achieve a common understanding of the rating criteria, the three observers independently assessed six pilot

Table 1 Rating of variables on magnetic resonance imaging of the lumbar spine

Variable	Description and rating categories
Modic changes, type [19–21]	Primary (the most extensive) and secondary signal intensity changes in the vertebral body marrow adjacent to the endplate rated as 0—no changes, type I—hypointense T1 signal and hyperintense T2 signal, type II—hyperintense T1 signal and iso- or slightly hyperintense T2 signal and type III—hypointense T1 signal and hypointense T2 signal
Modic changes, extent [11]	Maximal AP extent rated as <25%, 25−50% or >50% of AP endplate diameter on sagittal images Maximal CC extent rated as minimal (small dots), <25%, 25–50% or >50% of vertebral body height on sagittal images
Posterior HIZ [10, 14]	Area of high-signal intensity in the posterior annulus fibrosus that is brighter than the nucleus pulposus on T2-weighted images and is surrounded superiorly, inferiorly and anteriorly by the low-intensity (black) signal of the annulus fibrosus; rated as present or not present
Nucleus pulposus signal [22]	Nucleus visually rated as bright, grey, dark or black on sagittal T2-weighted images, using cerebrospinal fluid as intensity reference
Disc height [15, 23, 24]	Disc height narrowing visually rated by comparing to the disc above if it is normal, and otherwise and at L5/S1 based on experience, as 0—no, disc higher than disc above (if normal), 1—slight, disc as high as the disc above (if normal), 2—moderate, disc narrower than the disc above (if normal) or 3—severe, endplates almost in contact
Measured disc height decrease [25]	Distance measured in millimetres or pixels between the mid-inferior and the mid-superior disc borders on the mid-sagittal T2-weighted image, disc height calculated as a percentage of the nearest normal above disc height, and disc height decrease noted as <40% or ≥40%
Disc contour [19]	Rated as 0—normal, 1—bulge (base >1/2 of disc circumference) or 2—herniation (includes protrusion, extrusion and sequestration)
Facet arthropathy [10, 26]	Rated for worst side (right/left) on axial images or on sagittal images if axial images are lacking, as 0—normal, 1—mild (joint space narrowing or mild osteophyte), 2—moderate (sclerosis or moderate osteophyte) or 3—severe (marked osteophyte)

AP anteroposterior, *CC* craniocaudal, *HIZ* high-intensity zone

examinations from another study. Observers A and B then discussed ratings and criteria at a joint 2-h meeting. Observer C did not attend the meeting but compared ratings with observers A and B and discussed with the last author, who had attended.

Statistical analyses

All MRI findings were dichotomized into categories that reflected the inclusion criteria or that might be clinically relevant (see "Results" section). The prevalence of each type of dichotomised MRI finding was calculated at each rated level for each observer. As in similar studies [9, 11], only findings with a mean prevalence 10–90% across all observers at the rated level were further analysed, since very high or low prevalence can lead to very low agreement beyond chance, despite very high actual agreement [27]. Each finding was further analysed at each rated level. MRI indication for prosthesis (yes/no) was analysed separately at L4/L5 and L5/S1 and noted as present when the observer reported ≥40% disc height decrease and/or at least two of these three findings: Modic changes type I/II (superior and/or inferior to disc), posterior HIZ and dark/black nucleus pulposus. These retrospective reports were not used in the prospective trial.

Using STATA 10.0 (College Station, TX), unweighted overall kappa was computed for agreement between all observers with a 95% bias-corrected confidence interval based on bootstrapping with 1,000 repetitions. Unweighted kappa for pairwise interobserver agreement and for intra-observer agreement was calculated using SPSS 17.0 (SPSS, Chicago, IL). p values were computed for difference in the prevalence of findings across observers (fixed effects model, STATA 10.0). After Bonferroni adjustment for multiple comparisons, $p<0.002$ indicated statistical significance. Kappa was interpreted as: $k\leq0.20$, poor; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, good and 0.81–1.00, very good agreement beyond chance [28].

Sample size

For each comparison, if the true kappa is 0.60 and the prevalence 30%, 191 paired observations provide 80% power to give a significant result at the 5% level in a two-sided test of $k=0.40$ [27]. Three observers were used in order to improve the power in this study with a fixed sample size $n=170$.

Results

All observers rated all findings at L3–S1 in all 170 examinations, except for type of any Modic changes in the two examinations lacking T1 images. Observers A and B rated all findings twice in 126 cases for intraobserver

Table 2 Prevalence of findings in percent by reader

Finding	Reader A	Reader B	Reader C	p value[a]
Modic changes present				
L4/L5 sup to disc	31.8	28.8	44.7	<0.0001
L4/L5 inf to disc	37.1	26.5	52.9	<0.0001
L5/S1 sup to disc	73.5	69.4	80.6	0.0001
L5/S1 inf to disc	68.8	65.9	76.5	0.0001
AP extent of Modic changes >50% of endplate diameter[b]				
L4/L5 sup to disc	64.8	61.2	59.2	0.0006
L4/L5 inf to disc	54.0	57.8	52.2	0.0001[c]
L5/S1 sup to disc	85.6	79.7	84.7	<0.0001
L5/S1 inf to disc	86.3	82.1	87.7	<0.0001
CC extent of Modic changes >50% of vertebral body height[b]				
L4/L5 sup to disc	25.9	22.4	19.7	0.7391
L4/L5 inf to disc	17.5	13.3	15.6	0.2878
L5/S1 sup to disc	40.0	37.3	38.7	0.0665
L5/S1 inf to disc	12.8	8.9	23.1	<0.0001
Posterior HIZ present				
L4/L5	18.2	38.8	26.5	<0.0001
L5/S1	9.4	31.2	21.2	<0.0001
Nucleus pulposus signal dark/black				
L4/L5	54.1	55.9	42.4	<0.0001
L5/S1	72.4	71.2	57.6	<0.0001
Disc height judged severely reduced				
L4/L5	14.1	4.7	12.4	<0.0001
L5/S1	27.1	29.4	52.4	<0.0001
Measured ≥40% disc height decrease				
L4/L5	15.3	11.8	19.4	0.0014
L5/S1	57.1	54.4	65.9	0.0001
Disc contour abnormal (bulge/herniation)				
L4/L5	66.5	49.4	75.9	<0.0001
L5/S1	81.8	66.5	86.5	<0.0001
Facet arthropathy moderate/severe[d]				
L4/L5	14.1	5.9	14.1	0.0027
Disc prosthesis indicated[e]				
L4/L5	45.9	55.3	51.2	0.0053
L5/S1	79.4	82.4	82.9	0.3078

The data are based on magnetic resonance imaging in 170 patients

sup superior, *inf* inferior, *AP* anteroposterior, *CC* craniocaudal, *HIZ* high-intensity zone

[a] p value for difference in prevalence across observers (likelihood ratio test, fixed effects model)

[b] Modic changes extent is contingent on Modic changes being present

[c] p value based on generalized estimating equations, because convergence was not achieved using a fixed effects model

[d] Finding not analysed at L5/S1, since it had a mean prevalence <10% across observers at L5/S1

[e] Based on report of measured ≥40% disc height decrease and/or at least two of these three findings: Modic changes type I and/or II (superior and/or inferior to disc), posterior HIZ and nucleus pulposus signal dark/black

analysis. Due to a mean prevalence <10% in the n=170 sample, we did not further analyse any finding at L3/L4 or facet arthropathy at L5/S1.

Interobserver reliability

The prevalence at each rated level differed significantly (p< 0.002) but slightly across observers for most findings (Table 2). Observer C reported more Modic changes and twice as high prevalence as observer B at L4/L5 inferior to disc, i.e. at the upper endplate of L5 (52.9% vs. 26.5%, Table 2). The observers similarly often noted >50% CC extent of Modic changes, except at L5/S1 inferior to disc (Table 2). The prevalence at individual disc levels differed up to threefold between observers for posterior HIZ and for disc height judged severely reduced; it differed less for ≥40% measured disc height decrease, dark/black nucleus pulposus signal and abnormal disc contour (Table 2, Fig. 1). The difference in prevalence between observers was in a different direction for different findings (Table 2). Thus, the overall MRI indication for prosthesis did not differ significantly in prevalence across observers, neither at disc level L4/L5 nor at disc level L5/S1, but it tended to differ at L4/L5 (Table 2).

Overall agreement was moderate or good (k=0.56–0.77) for presence and extent of Modic changes, but only fair (k= 0.40) for inferior CC extent at L5/S1 (Table 3), which had a low mean prevalence across observers (14.7%). Regarding HIZ, overall agreement was moderate but better at L4/L5 than L5/S1 (k=0.58 vs. 0.46, Table 3). Overall agreement was moderate or good (k=0.50–0.72) for dark/black nucleus pulposus signal, severely reduced disc height, ≥40% measured disc height decrease and abnormal disc contour, and fair (k=0.24) for moderate/severe facet arthropathy at L4/L5 (Table 3), which had a mean prevalence across observers of 11.4%. The MRI indication for disc prosthesis showed good overall agreement both at L4/L5 (k=0.70) and at L5/S1 (k=0.66).

Pairwise agreement ranged from fair to very good. It was fair in one pair at L5/S1 for inferior AP and CC extent of Modic changes, superior AP extent, posterior HIZ and disc contour, and in all pairs for facet arthropathy at L4/L5. It was otherwise moderate to very good (Table 3).

Intraobserver reliability

Intraobserver agreement was good or very good (k=0.61–1.00) except in one observer at L5/S1 for inferior AP and CC extent of Modic changes (k=0.38–0.55) and for HIZ (k= 0.60, Table 4). It was mostly very good (k=0.67–0.87) for the indication for prosthesis (Table 4).

Discussion

In this study, interobserver agreement was generally moderate or good for findings included in the indication for disc prosthesis (Modic changes, HIZ, dark/black nucleus pulposus, ≥40% disc height decrease) but only fair for facet arthropathy. Intraobserver agreement was mostly good or very good.

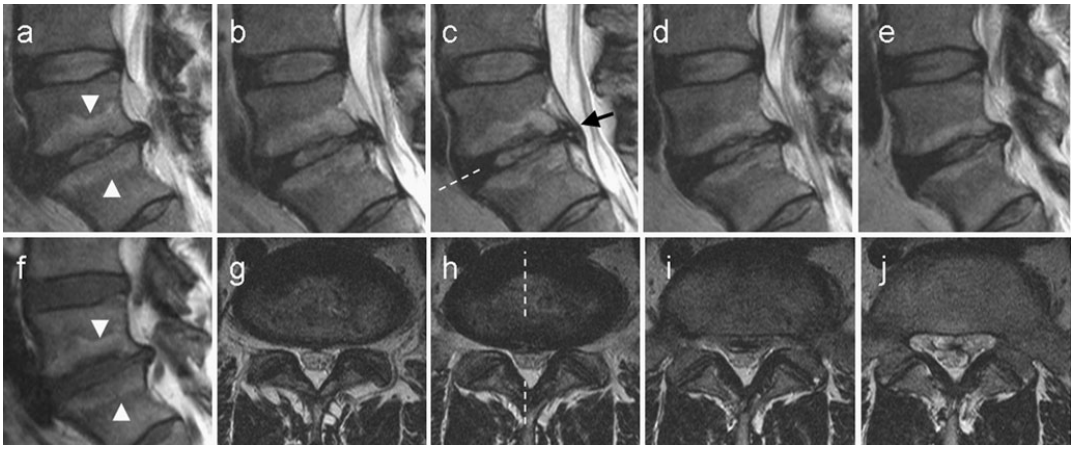

Fig. 1 Magnetic resonance imaging of one patient; sagittal T2-weighted images (**a**–**e**) shown in the order of patient's left to right, sagittal T1-weighted image (**f**) corresponding to T2-weighted image in **a**, and axial T2-weighted images (**g**–**j**) shown from cranially to caudally. Image plane shown in **c** is marked on **h** and vice versa (*broken lines*). At L5/S1, all observers agreed on Modic changes primary type II (**a**, **f**; *arrow heads*), grey nucleus pulposus on T2-weighted images (**a**–**e**), ≥40% measured disc height decrease compared to the normal disc above, disc herniation, and slight facet arthropathy (**h**–**j**) but not on posterior high-intensity zone (**c**, *arrow*) or severely reduced disc height judged subjectively

Table 3 Interobserver agreement measured by using the kappa statistic

Finding	Readers A and B	Readers A and C	Readers B and C	Overall kappa (95% confidence interval)
Modic changes present				
L4/L5 sup to disc	0.90	0.68	0.64	0.73 (0.65, 0.82)
L4/L5 inf to disc	0.76	0.55	0.44	0.56 (0.47, 0.66)
L5/S1 sup to disc	0.81	0.60	0.55	0.67 (0.57, 0.76)
L5/S1 inf to disc	0.85	0.66	0.63	0.72 (0.63, 0.80)
AP extent of Modic changes >50% of endplate diameter[a]				
L4/L5 sup to disc	0.77	0.76	0.61	0.75 (0.57, 0.88)
L4/L5 inf to disc	0.81	0.74	0.52	0.72 (0.57, 0.88)
L5/S1 sup to disc	0.60	0.69	0.35	0.62 (0.46. 0.75)
L5/S1 inf to disc	0.55	0.52	0.23	0.56 (0.37, 0.70)
CC extent of Modic changes >50% of vertebral body height[a]				
L4/L5 sup to disc	0.78	0.65	0.49	0.62 (0.39, 0.80)
L4/L5 inf to disc	0.83	0.64	0.66	0.77 (0.59, 0.92)
L5/S1 sup to disc	0.73	0.74	0.55	0.67 (0.55, 0.77)
L5/S1 inf to disc	0.51	0.52	0.28	0.40 (0.23, 0.56)
Posterior HIZ present				
L4/L5	0.49	0.66	0.62	0.58 (0.46, 0.68)
L5/S1	0.37	0.47	0.56	0.46 (0.34, 0.58)
Nucleus pulposus signal dark/black				
L4/L5	0.73	0.68	0.69	0.69 (0.61, 0.78)
L5/S1	0.59	0.53	0.54	0.55 (0.45, 0.64)
Disc height judged severely reduced				
L4/L5	0.46	0.82	0.52	0.62 (0.46, 0.77)
L5/S1	0.80	0.51	0.53	0.58 (0.48, 0.68)
Measured ≥40% disc height decrease				
L4/L5	0.70	0.73	0.62	0.69 (0.53, 0.80)
L5/S1	0.74	0.72	0.70	0.72 (0.64, 0.79)
Disc contour abnormal (bulge/herniation)				
L4/L5	0.64	0.75	0.47	0.60 (0.49, 0.70)
L5/S1	0.55	0.69	0.35	0.51 (0.38, 0.65)
Facet arthropathy moderate/severe[b]				
L4/L5	0.29	0.22	0.23	0.24 (0.06, 0.42)
Disc prosthesis indicated[c]				
L4/L5	0.72	0.68	0.71	0.70 (0.62, 0.79)
L5/S1	0.68	0.69	0.61	0.66 (0.54, 0.76)

The data are unweighted kappa values based on magnetic resonance imaging in 170 patients.

sup superior, *inf* inferior, *AP* anteroposterior, *CC* craniocaudal, *HIZ* high-intensity zone

[a]Kappa values are based on a subsample with Modic changes present according to both observers or according to all observers for overall kappa

[b]Kappa value at L5/S1 not given, since the finding had mean prevalence <10% at L5/S1

[c]Based on report of measured ≥40% disc height decrease and/or at least two of these three findings: Modic changes type I and/or II (superior and/or inferior to disc), posterior HIZ and nucleus pulposus signal dark/black

Modic changes, HIZ and severely reduced disc height judged subjectively differed up to two- or threefold in prevalence between observers at individual disc levels. The overall MRI indication for disc prosthesis showed more similar prevalence across observers and good interobserver and intraobserver agreement both at L4/L5 and at L5/S1.

Strengths and limitations

The strengths of our study included the use of three observers, a large sample (n=170) in the interobserver analysis, the analysis of separate disc levels and the testing of disagreement on prevalence. Such disagreement (bias) cannot be assessed by means of the kappa coefficient; it reduces expected agreement by chance and actually increases the kappa values slightly [27]. Disagreement between observers on the prevalence of a finding shows that their ratings of the finding differ systematically. Systematic differences in the interpretation of important findings should be identified by appropriate methods and addressed to improve the reliability.

The observers used well-defined MRI rating criteria, but they knew the patients were accepted for disc prosthesis surgery due to localised degeneration. How this may have

Table 4 Intraobserver agreement measured by using the kappa statistic

Finding	Reader A	Reader B
Modic changes present		
L4/L5 sup to disc	0.88	0.89
L4/L5 inf to disc	0.82	0.80
L5/S1 sup to disc	0.90	0.87
L5/S1 inf to disc	0.83	0.95
AP extent of Modic changes >50% of endplate diameter[a]		
L4/L5 sup to disc	0.89	0.88
L4/L5 inf to disc	0.86	0.85
L5/S1 sup to disc	0.79	0.61
L5/S1 inf to disc	–[b]	0.38
CC extent of Modic changes >50% of vertebral body height[a]		
L4/L5 sup to disc	0.94	0.87
L4/L5 inf to disc	0.86	1.00
L5/S1 sup to disc	0.70	0.67
L5/S1 inf to disc	0.55	–[b]
Posterior HIZ present		
L4/L5	0.85	0.81
L5/S1	0.88	0.60
Nucleus pulposus signal dark/black		
L4/L5	0.86	0.76
L5/S1	0.83	0.69
Disc height judged severely reduced		
L4/L5	0.74	–[b]
L5/S1	0.77	0.75
Measured ≥40% disc height decrease		
L4/L5	0.77	0.64
L5/S1	0.81	0.67
Disc contour abnormal (bulge/herniation)		
L4/L5	0.83	0.76
L5/S1	0.81	0.76
Facet arthropathy moderate/severe		
L4/L5	0.67	–[b]
Disc prosthesis indicated[c]		
L4/L5	0.87	0.82
L5/S1	0.85	0.67

The data are unweighted kappa values based on magnetic resonance imaging in 126 patients

sup superior, *inf* inferior, *AP* anteroposterior, *CC* craniocaudal, *HIZ* high-intensity zone

[a] Kappa values are based on a subsample with Modic changes present in both readings

[b] Kappa value not given because the finding had a mean prevalence >90% (AP extent of Modic changes) or <10% (other findings) in the first and second readings

[c] Based on report of measured ≥40% disc height decrease and/or at least two of these three findings: Modic changes type I and/or II (superior and/or inferior to disc), posterior HIZ and nucleus pulposus signal dark/black

affected their MRI ratings and agreement is not clear. The three radiologists came from different institutions, were not trained together and rated a range of findings on images obtained using different scanners and protocols. The often moderate reliability found in our study may therefore be representative for radiological subspecialty spine imaging practices.

Our results for patients accepted for disc prosthesis surgery should apply equally well to similar patients accepted for surgery with lumbar fusion. These reliability results provide a basis for further research on the role of MRI findings within both of these groups. Some of the results may also have a wider relevance. However, the reliability of the MRI indication for disc prosthesis surgery must be confirmed in chronic LBP patients not yet selected for surgery. Such patients may have a broader spectrum of MRI findings, causing more disagreement.

Discussion of results

We found clear differences in prevalence between observers for Modic changes, HIZ and subjectively rated disc height, and smaller differences for nucleus signal and abnormal disc contour, whereas Carrino et al. [10] found differences in frequency distributions between trained experts for disc degeneration (p=0.055, Wald test) and facet arthropathy (p=0.006) but not for Modic changes (p=0.52) or HIZ (p=0.22). No further comparable data exist. Lurie et al. [16] found similar frequencies across readers for bulges and normal discs combined.

It is noteworthy that the difference in prevalence between observers was in a different direction for different findings and did not add up to an even larger disagreement on the MRI indication for prosthesis. For example, observer B tended to report a lower prevalence of Modic changes and ≥40% disc height decrease than observer C but a higher prevalence of HIZ and dark/black nucleus signal and thus a more similar prevalence of the overall MRI indication (Table 2).

Disagreement on prevalence might be due to differences in interpretation and the use of rating criteria. It might also be due to differences in the observers' response bias, i.e. their tendency to prefer one or another response category (to rate up or down, particularly when in doubt), independently of the characteristics of the object [29]. Improved rating criteria might perhaps lower the number of ambiguous cases leading to differences in interpretation or response bias.

Our kappa values for interobserver and intraobserver agreement were generally similar or higher than in some prior studies for Modic changes [10], HIZ [9, 10, 12, 13], nucleus pulposus signal and disc height combined [9, 10, 15] and abnormal disc contour [9] but were similar [18] or

lower [10, 17] for facet arthropathy. This may be partly due to non-standardized images and low prevalence of moderate/severe facet arthropathy in our sample (11.4% at L4/L5). In three studies based on standardized MRI of 40-year olds from the normal population, kappa values were slightly higher for Modic changes [11], HIZ [15] and abnormal disc contour [12]. The observers in one of these studies had read 50 pilot examinations in consensus [15]. Overall, lumbar MRI findings show mostly moderate interobserver agreement.

There is no firm rule for when the reliability of a finding is adequate, and the use of multiple readers, e.g. in a study, might improve the rating of a finding [30]. Yet, we suggest that kappa ≤0.40 for interobserver agreement should lead to an assessment of how to improve the reliability. We found pairwise kappa ≤0.40 in one observer pair at L5/S1 for inferior extent of Modic changes, disc contour and HIZ. Agreement on HIZ might be improved by looking more closely at both axial and sagittal images and at the signal intensity compared to nucleus. It is also clear that better reliability is needed for facet arthropathy. This finding may be easier to rate on computed tomography (CT) [17, 18].

The clinical relevance of the studied MRI findings is not clear. Systematic reviews indicate that Modic changes are not yet documented to affect treatment outcome [31], that disc findings have only a weak and no clinically meaningful relation to LBP [32] and that there is no test that could identify facet joint arthropathy as source of pain [33]. Further studies are needed to clarify the relevance of such localised MRI findings for surgery with disc prosthesis.

Conclusions

Present state of the art in lumbar imaging shows mostly moderate interobserver agreement [9, 10]. In this study, the agreement was moderate to good for Modic and disc findings and only fair for facet arthropathy. Specific causes of disagreement and strategies to reduce it should be explored. The high reliability of the proposed MRI indication for prosthesis must be confirmed in unselected chronic LBP patients. Further studies are needed to assess the clinical relevance of these MRI findings in candidates for surgery with disc prosthesis or lumbar fusion.

Acknowledgements We would like to thank the patients who participated in this study. The study received financial support from the Haakon and Sigrun Ødegaard's fund at the Norwegian Society of Radiology, the South Eastern Norway Regional Health Authority and the Norwegian ExtraFoundation for Health and Rehabilitation through the Norwegian Back Pain Association.

Conflict of interest We declare that we have no conflict of interest.

References

1. Berg S, Tullberg T, Branth B, Olerud C, Tropp H (2009) Total disc replacement compared to lumbar fusion: a randomised controlled trial with 2-year follow-up. Eur Spine J 18:1512–1519
2. van den Eerenbeemt KD, Ostelo RW, van Royen BJ, Peul WC, van Tulder MW (2010) Total disc replacement surgery for symptomatic degenerative lumbar disc disease: a systematic review of the literature. Eur Spine J 8:1262–1280
3. Hellum C, Johnsen LG, Storheim K, Nygaard ØP, Brox JI, Rossvoll I, Rø M, Sandvik L, Grundnes O, and the Norwegian Spine Study Group (2011) Surgery with disc prosthesis versus rehabilitation in patients with low back pain and degenerative disc: two year follow-up of randomised study. BMJ 342:d2786. doi:10.1136/bmj.d2786
4. Harrop JS, Youssef JA, Maltenfort M, Vorwald P, Jabbour P, Bono CM, Goldfarb N, Vaccaro AR, Hilibrand AS (2008) Lumbar adjacent segment degeneration and disease after arthrodesis and total disc arthroplasty. Spine (Phila Pa 1976) 33:1701–1707
5. Park CK, Ryu KS, Jee WH (2008) Degenerative changes of discs and facet joints in lumbar total disc replacement using ProDisc II: minimum two-year follow-up. Spine (Phila Pa 1976) 33:1755–1761
6. Feinstein AR (1983) An additional basic science for clinical medicine: IV. The development of clinimetrics. Ann Intern Med 99:843–848
7. Jarvik JG, Deyo RA (2009) Moderate versus mediocre: the reliability of spine MR data interpretations. Radiology 250:15–17
8. Peterson CK, Gatterman B, Carter JC, Humphreys BK, Weibel A (2007) Inter- and intraexaminer reliability in identifying and classifying degenerative marrow (Modic) changes on lumbar spine magnetic resonance scans. J Manipulative Physiol Ther 30:85–90
9. Arana E, Royuela A, Kovacs FM, Estremera A, Sarasíbar H, Amengual G, Galarraga I, Martínez C, Muriel A, Abraira V, Gil Del Real MT, Zamora J, Campillo C (2010) Lumbar spine: agreement in the interpretation of 1.5-T MR images by using the Nordic Modic Consensus Group classification form. Radiology 254:809–817
10. Carrino JA, Lurie JD, Tosteson AN, Tosteson TD, Carragee EJ, Kaiser J, Grove MR, Blood E, Pearson LH, Weinstein JN, Herzog R (2009) Lumbar spine: reliability of MR imaging findings. Radiology 250:161–170
11. Jensen TS, Sorensen JS, Kjaer P (2007) Intra- and interobserver reproducibility of vertebral endplate signal (modic) changes in the lumbar spine: the Nordic Modic Consensus Group classification. Acta Radiol 48:748–754
12. Kovacs FM, Royuela A, Jensen TS, Estremera A, Amengual G, Muriel A, Galarraga I, Martínez C, Arana E, Sarasíbar H, Salgado RM, Abraira V, López O, Campillo C, del Real MT, Zamora J (2009) Agreement in the interpretation of magnetic resonance images of the lumbar spine. Acta Radiol 50:497–506
13. Smith BM, Hurwitz EL, Solsberg D, Rubinstein D, Corenman DS, Dwyer AP, Kleiner J (1998) Interobserver reliability of detecting lumbar intervertebral disc high-intensity zone on magnetic resonance imaging and association of high-intensity zone with pain and anular disruption. Spine 23:2074–2080
14. Aprill C, Bogduk N (1992) High-intensity zone: a diagnostic sign of painful lumbar disc on magnetic resonance imaging. Br J Radiol 65:361–369

15. Solgaard Sorensen J, Kjaer P, Jensen ST, Andersen P (2006) Low-field magnetic resonance imaging of the lumbar spine: reliability of qualitative evaluation of disc and muscle parameters. Acta Radiol 47:947–953
16. Lurie JD, Tosteson AN, Tosteson TD, Carragee E, Carrino JA, Kaiser J, Sequeiros RT, Lecomte AR, Grove MR, Blood EA, Pearson LH, Herzog R, Weinstein JN (2008) Reliability of magnetic resonance imaging readings for lumbar disc herniation in the Spine Patient Outcomes Research Trial (SPORT). Spine (Phila Pa 1976) 33:991–998
17. Weishaupt D, Zanetti M, Boos N, Hodler J (1999) MR imaging and CT in osteoarthritis of the lumbar facet joints. Skeletal Radiol 28:215–219
18. Stieber J, Quirno M, Cunningham M, Errico TJ, Bendo JA (2009) The reliability of computed tomography and magnetic resonance imaging grading of lumbar facet arthropathy in total disc replacement patients. Spine (Phila Pa 1976) 34:E833–E840
19. Fardon DF (2001) Nomenclature and classification of lumbar disc pathology. Spine (Phila Pa 1976) 26:461–462
20. Modic MT, Ross JS (2007) Lumbar degenerative disk disease. Radiology 245:43–61
21. Modic MT, Steinberg PM, Ross JS, Masaryk TJ, Carter JR (1988) Degenerative disk disease: assessment of changes in vertebral body marrow with MR imaging. Radiology 166:193–199
22. Luoma K, Riihimaki H, Luukkonen R, Raininko R, Viikari-Juntura E, Lamminen A (2000) Low back pain in relation to lumbar disc degeneration. Spine (Phila Pa 1976) 25:487–492
23. Raininko R, Manninen H, Battie MC, Gibbons LE, Gill K, Fisher LD (1995) Observer variability in the assessment of disc degeneration on magnetic resonance images of the lumbar and thoracic spine. Spine (Phila Pa 1976) 20:1029–1035
24. Videman T, Battie MC, Gibbons LE, Maravilla K, Manninen H, Kaprio J (2003) Associations between back pain history and lumbar MRI findings. Spine (Phila Pa 1976) 28:582–588
25. Masharawi Y, Kjaer P, Bendix T, Manniche C, Wedderkopp N, Sorensen JS, Peled N, Jensen TS (2008) The reproducibility of quantitative measurements in lumbar magnetic resonance imaging of children from the general population. Spine (Phila Pa 1976) 33:2094–2100
26. Fujiwara A, Tamai K, Yamato M, An HS, Yoshida H, Saotome K, Kurihashi A (1999) The relationship between facet joint osteoarthritis and disc degeneration of the lumbar spine: an MRI study. Eur Spine J 8:396–401
27. Sim J, Wright CC (2005) The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 85:257–268
28. Altman D (1991) Practical statistics for medical research. Chapman & Hall, New York
29. Ker M (1991) Issues in the use of kappa. Invest Radiol 126:78–83
30. Robinson PJ (1997) Radiology's Achilles' heel: error and variation in the interpretation of the Rontgen image. Br J Radiol 70:1085–1098
31. Jensen RK, Leboeuf-Yde C (2011) Is the presence of Modic changes associated with the outcomes of different treatments? A systematic critical review. BMC Musculoskelet Disord 12:183
32. Endean A, Palmer KT, Coggon D (2011) Potential of magnetic resonance imaging findings to refine case definition for mechanical low back pain in epidemiological studies: a systematic review. Spine (Phila Pa 1976) 36:160–169
33. Hancock MJ, Maher CG, Latimer J, Spindler MF, McAuley JH, Laslett M, Bogduk N (2007) Systematic review of tests to identify the disc, SIJ or facet joint as the source of low back pain. Eur Spine J 16:1539–1350